1-1-2015

The Australian Medical Schools Assessment Collaboration: benchmarking the preclinical performance of medical students

Deborah A. O'Mara
Sydney Medical School

Ben Canny
Monash University

Imogene P. Rothnie
Sydney Medical School

Ian Wilson
University of Wollongong, ianwil@uow.edu.au

John Barnard
Excel Psychological and Educational Consultancy

See next page for additional authors

The Australian Medical Schools Assessment Collaboration: benchmarking the preclinical performance of medical students

Abstract

Objectives: To report the level of participation of medical schools in the Australian Medical Schools Assessment Collaboration (AMSAC); and to measure differences in student performance related to medical school characteristics and implementation methods. Design: Retrospective analysis of data using the Rasch statistical model to correct for missing data and variability in item difficulty. Linear model analysis of variance was used to assess differences in student performance. Setting and participants: 6401 preclinical students from 13 medical schools that participated in AMSAC from 2011 to 2013. Main outcome measures: Rasch estimates of preclinical basic and clinical science knowledge. Results: Representation of Australian medical schools and students in AMSAC more than doubled between 2009 and 2013. In 2013 it included 12 of 19 medical schools and 68% of medical students. Graduate-entry students scored higher than students entering straight from school. Students at large schools scored higher than students at small schools. Although the significance level was high ($P < 0.001$), the main effect sizes were small (4.5% and 2.3%, respectively). The time allowed per multiple choice question was not significantly associated with student performance. The effect on performance of multiple assessments compared with the test items as part of a single end-of-year examination was negligible. The variables investigated explain only 12% of the total variation in student performance. Conclusions: An increasing number of medical schools are participating in AMSAC to monitor student performance in preclinical sciences against an external benchmark. Medical school characteristics account for only a small part of overall variation in student performance. Student performance was not affected by the different methods of administering test items.

Disciplines

Medicine and Health Sciences | Social and Behavioral Sciences

Publication Details

O'Mara, D. A., Canny, B., Rothnie, I. P., Wilson, I. G., Barnard, J. & Davies, L. (2015). The Australian Medical Schools Assessment Collaboration: benchmarking the preclinical performance of medical students. Medical Journal of Australia, 202 (2), 95-98.

Authors

Deborah A. O'Mara, Ben Canny, Imogene P. Rothnie, Ian Wilson, John Barnard, and Llewelyn Davies

The Australian Medical Schools Assessment Collaboration: benchmarking the preclinical performance of medical students

Deborah A O'Mara BA(Hons), DipEd, PhD1

Ben J Canny BMedSc(Hons), MB BS, PhD2

Imogene P Rothnie BA BSc, PostGradDipPsychol, MEd1

Ian G Wilson MAssess&Eval, PhD, FRACGP3

John Barnard MSc, MEd, PhD4

Llewelyn Davies MB BS, MD, FRACP1,5

1 Sydney Medical School, Sydney, NSW.

2 Monash University, Melbourne, VIC.

3 University of Wollongong, Wollongong, NSW.

4 EPEC (Excel Psychological and Educational Consultancy),Melbourne, VIC.

5 Royal Prince Alfred Hospital, Sydney, NSW.deborah.omara@sydney.edu.au

Objectives: The Australian Medical Schools Assessment Collaboration (AMSAC) was established in 2008 by 6 medical schools to enable monitoring of their students' performance with those of other schools.

Design, setting and participants: From 2011-2013 the benchmark involved 8729 students in pre-clinical years from 14 Australian medical schools. A set of up to 50 items are embedded each year in existing summative assessments. Data are analysed using the Rasch statistical model to correct for missing data and variability in the difficulty of items used by individual schools. Differences in student performance have been investigated by characteristics of medical schools and implementation methods using a linear model analysis of variance.

Main outcome measures: Knowledge and application of the basic clinical sciences.

Results: Graduate entry students scored higher than students entering straight from school. Students from large schools scored higher than students from small schools. Although the significance level was high ($p<0.001$), the main effect sizes were small; 4.5% and 2%, respectively. The time allowed per MCQ had no significant effect. Students who were tested over multiple assessments did not score significantly higher than those presented with the test items as part of a single end of year examination. The variables investigated explain only 12% of the total variation in student performance.

Conclusion: AMSAC provides an opportunity to monitor student performance in pre-clinical science against an external benchmark. Differences in performance by medical school type account for a small variation in student performance. The main source of variation is individual student ability.

1 INTRODUCTION

There was a rapid expansion of medical schools in Australia in the last decade with the addition of 7 new schools at Bond University, Deakin University, Griffith University, James Cook University, The University of Notre Dame, University of Western Sydney and University of Wollongong. Australia now has 19 medical schools. The various schools have a number of differentiating characteristics including entry point, curriculum, course duration and teaching methods. Some schools identify themselves as trying to achieve specific graduate attributes such as producing future specialists, rural health practitioners or medical researchers while others are more general in their aims.

Medical schools need a process that allows them to measure changes in their students' and graduates' performance relative to other medical schools so that they can evaluate changes to their selection criteria, curriculum and teaching methods. The Australian Medical Schools Assessment Collaboration (AMSAC) was established by a group of six medical schools in 2008. It aims to provide a means for medical schools in Australia to share assessment items and performance statistics within a professional and anonymous framework. The formation principles guard against the construction of league tables. Currently 13 Australian medical schools are part of the AMSAC collaboration (Appendix 1 online at mja.com.au).

One of the educational benefits of the collaboration has been the collegial activity generated by item submission and review. Institutional variation in the quality of assessments has been noted overseas (1, 2) but there have been few reports on variation in medical school assessments in Australia (3). While assessment collaborations exist for item sharing there has not been any previous sharing of Australian student performance data. The collaboration assesses "pre-clinical" medical education in the sense that schools generally administer the items at the point where students make the transition from predominantly campus teaching to mainly clinical settings. This division is more clear-cut in some schools than others.

We report the implementation and educational benefits of the AMSAC collaboration as well as differences in student performance related to medical school characteristics and test administration methods.

2 METHODS

AMSAC cohort

The representation of Australian medical schools and students in AMSAC has more than doubled since it began in 2008. In 2013 it included 12 of 19 medical schools and 68% of medical students (Box 1). Since 2010 most AMSAC collaborators (84%) have embedded the items in second year examinations irrespective of the total length of their programs. Year two enrolments for schools outside the collaboration have been used to calculate the “equivalent year” population base (Box 1).

Fitting questions into all school curricula is challenging. Although the initial collaboration involved seven medical schools, one was unable to field questions in 2009 and another in 2010. Two medical schools participated for two years and withdrew due to difficulty in matching the agreed blueprint within a single cohort; one has recently rejoined.

The proportion of graduate entry schools has grown from half of the participating schools in 2009 to two-thirds in 2013. The proportion of students who participate in AMSAC and attend graduate entry school has been relatively stable; 69% in 2009, 62% in 2011 and 67% in 2013. Similarly, the proportion of AMSAC schools that could be classified as large (intake of 150 or more) ranged from 50% in 2009 to 56% in 2011 and 50% in 2013 with the corresponding student representation being 61% in 2009, 74% in 2011 and 73% in 2013.

In terms of representing the Australian medical school population, in 2013, AMSAC included six of the ten small schools, six of the nine large schools, eight of the eleven graduate-entry schools and four of the eight school-leaver entry schools.

AMSAC Assessment Generation

The project creates an agreed set of 50 items for participating schools to include in summative assessments. A multiple choice question (MCQ) Type A format is used with one correct answer and four distractors. This is the most widely used written item type in assessment of basic and clinical sciences (4, 5) and, if well designed, assess reasoning as well as factual recall (6). Although some Australian schools use four option MCQ questions, most use the five option format standardised by the United States National Board of Medical Examiners (7). Although studies suggest that varying the number of options between three and five has little influence on item performance (8, 9) the collaboration has used the five-option convention because of its widespread acceptance.

The items are mapped to a blueprint (Appendix 2 online at mja.com.au) that covers two broad basic science domains; “function” and “structure” and are managed through the Sydney Medical School’s assessment database (10). All participating schools contribute items and these are reviewed at an annual collaborators’ meeting. A short list of 60 items is circulated and any items nominated by multiple schools as problematic are eliminated to produce a final set of 50. Approximately half the items are anchor items that have been used previously and have performed well.

AMSAC implementation

The AMSAC items are delivered to a single student cohort in the collaborating schools as part of one or more summative assessments over a calendar year. Schools vary in terms of the number of items they include in their assessments (Appendix 3 online at mja.com.au) because item relevance is based on assessment timing and curriculum delivery. This is in part due to variable use of semester, unit and end of year examinations.

The time allowed for each item in an assessment varies between collaboration members; from a low of 60 seconds to a high of 120 seconds. In 2009 and 2010 students from schools allowing 90 or more seconds performed significantly better than those allowing 60 seconds and one medical school changed their assessment procedures as a result of this finding (11).

Statistical analysis

Each year, the performance of individual students on the AMSAC items is collated and analysed by an independent consultancy (EPEC), allowing schools to preserve their anonymity through the use of confidential identifiers. The data are analysed using the Rasch model which, unlike classical test theory, accounts for missing data in estimates of item difficulty and student performance and enables valid comparisons to be made across the cohort, irrespective of whether all 50 items are administered. Rasch analysis has been applied widely in medical education assessments (12, 13). The Rasch Measure score (Winsteps Version 3.80.1) for each student was used to investigate differences by type of medical school and implementation variations. Although Rasch estimates can be derived for both domains (structure and function), some schools implemented too few questions to provide a reliable basis for analysis of the individual domains.

Statistical analysis was subsequently undertaken using a general linear model analyses of variance (GLM ANOVA) in IBM SPSS Version 21(14). Five independent variables (year of administration, entry requirement, school size, number of assessments and time per item) were analysed. Significance was defined at $p<0.01$.Results

Rasch analysis found the item set to be psychometrically sound in each year with a good fit of items to the model (Appendix 2 online at mja.com.au) and ability estimates for student performance are robust. The classical reliability of the 2013 AMSAC question cohort was 0.85 on the KR-20 index. The average student performance varied slightly across the five implementations. The Rasch model sets the mean question difficulty at zero. Because most students correctly answered more than half of the questions the mean of the measure of student performance is greater than zero. The apparent variation in student performance in the early iterations was due to variation in item difficulty as well as the smaller number of schools participating (Appendix 3 online at mja.com.au), so the ANOVAs have been carried out only on 2011-2013 data. As the number of reliable marker questions has increased over the years the performance of the question set has stabilised. A difference of half a logit in student Rasch scores is considered significant for high stakes assessments(15).

On ANOVA there was a significant difference ($p<0.001$) for three of the four independent variables with entry point being the most significant main effect followed by size of school and number of assessments (Appendix 4 online at mja.com.au). The effect of time allowed per MCQ was not significant ($p=0.681$). The mean and standard deviation of student performance for each interaction clarifies the direction of the significant difference (Box 4).

Multiple comparisons for Year of AMSAC implementation were the same for each ANOVA with the Scheffe test identifying the mean score for 2013 as being significantly different to both 2011 (MD=-0.22 p<0.001) and 2012 (MD=-0.18 p<0.001). The mean scores for 2011 and 2012 were not significantly different (MD=0.04 p=>0.05). It should be noted that although the difference in the means was statistically significant this difference was less than one quarter of a logit and therefore not substantive. There was a large overlap between the inter-quartile ranges of the distribution of student performance measured with the Rasch model each year (Box 3).

Students from larger schools (>150 students) perform better than students from smaller schools. The degree of difference has increased over time and is significant in recent years (Appendix 4 online at mja.com.au), however the difference attributable to school size explains only 2% of variation and the interaction with year of implementation a further 1%. The difference between means for school size increased over time, being largest in 2013 (0.38) but still less than 0.5 of a logit. By ANOVA the graduate entry students performed significantly better than those entering medicine directly from school (Appendix 4 online at mja.com.au). The proportion of variance attributable to entry type is 4.5%. The pooled interaction with year of implementation varied over time (Box 4).

In the first year of the implementation the time allowed per item was significant with students allowed 60 seconds doing less well than those allowed 90 seconds per item(1). In subsequent years the time allowed was not significant (p=0.681). The effect size for delivery in a single year-end assessment or in multiple exams was too low to be meaningful (partial eta^2=0.003) although the p value for the test was significant (Appendix 4 online at mja.com.au). The interaction of number of assessments and year of implementation explains a low proportion of variance. The frequency of summative assessments has increased each year of the collaboration and therefore some medical schools are re-classified over time (Box 4).

There was no inter-dependency between the independent variables used in this study. The largest association was between size of medical school and whether AMSAC was used in more than one assessment (r=0.44), indicating that larger schools tend to employ more assessments. The multiple regression with all five independent variables explained only

12% of total variance ($R^{2=}$ 0.12). The standardised *Beta* weights were 0.28 for entry type, 0.26 for size of school, 0.14 for time per MCQ item, -0.10 for AMSAC year and -0.07 for number of assessments -0.07. Thus 88% of variation in medical student performance is due to factors outside the model, most likely individual student ability.

3 Discussion

The AMSAC project demonstrates the viability of linking assessments across medical schools with as few as 25 common items enabling reliable inter-school comparisons. Individual medical schools can use the AMSAC data to assess the effect of changes in relevant curriculum or entry requirements and to evaluate the need for change.

AMSAC is a broad collaboration of medical schools that vary in size, selection criteria, course duration and syllabus content. The overall sampling of Australian medical school students has grown from one third of the pre-clinical cohort in 2009 to over two thirds in 2013 and will include 14 schools in 2015. Two schools have left the collaboration due to difficulty with curriculum mapping, though one has recently rejoined.

A key outcome of this process has been the collegiate interaction of the medical school representatives. The schools involved have acquired a better understanding of the structure and content of other schools' syllabuses. The increase in the quality and stability of the items during the project is a reflection of the broad engagement of course leaders in the question collection and review process, which has also been found in overseas collaborations (16).

Rasch analysis copes well with missing data and allows valid comparisons between schools using less than the full set of test items and has enabled valid comparisons to be made by medical school and implementation methods. The slightly better performance of graduate entry students in the early years of medical school may reflect increased maturity and previous success but likely also reflects the substantial percentage with a medical science degree (27% of entrants Sydney Medical School 2011-2013). It would be surprising if an additional three years of study in the medical sciences did not confer any advantage in a medical science examination. A previous study from Melbourne Medical School found a similar small advantage for graduate entry students (17).

The small but significant difference in performance by size of school probably reflects the fact that all of the new Australian medical schools are small. These new schools are unlikely to have the depth of resources of the established schools in pre-clinical sciences and are also in a stage of course stabilization as their early cohorts reach graduation. The overwhelming source of variation in student performance on the AMSAC items is due to other factors, most likely individual student ability. It is acknowledged that other benefits that are derived from the quality of student-teacher interaction at small schools is not assessed in the current AMSAC outcome measure.

AMSAC provides an opportunity for comparison of assessment strategies across schools and over time there has been an increase in schools using multiple assessments as opposed to a single end of year exam. This pattern is particularly strong amongst the larger schools perhaps reflecting better resourcing. Student performance is not measurably better in multiple small assessments.

The project allows participating schools some ability to compare their students' knowledge base and reasoning skills with those of other schools and a national average. This project does not address other graduate attributes such as clinical decision making and only looks at mid-course performance but other national collaborations are in process to examine clinical skills and knowledge and reasoning in the pre-graduation phase.

The project has demonstrated that medical schools can collaborate on a benchmarking process without the need for external regulation. Early fears about misuse of the data to create league tables or damage to unique syllabus content have not been realised. The AMSAC project is a model for national collaboration between medical schools to meet government and community demands for accountability without loss of school autonomy.

References
References should be in Vancouver style and should **not** appear as endnotes. References to material on the Internet should include the organisation, the page title, the article title and the author (if there is one) as well as the URL and the month the page was visited (see examples here).

1. Reichert TG. Assessing The Use of High Quality Multiple Choice Exam Questions in Undergraduate Nursing Education: Are Educators Making the Grade? 2011.

2. Jozefowicz RF, Koeppen BM, Case S, et al; The quality of in-house medical school examinations. Acad Med. 2002;77(2):156-61.

3. Edwards D, Wilkinson D, Canny BJ, et al;. Developing outcomes assessments for collaborative, cross-institutional benchmarking: Progress of the Australian Medical Assessment Collaboration. Med Teach. 2014;36(2):139-47.

4. Schuwirth LW, van der Vleuten CP. General overview of the theories used in assessment: AMEE Guide No. 57. Med Teach. 2011;33(10):783-97.

5. Norcini J, Swanson D, Grosso L, Webster G. Reliability, validity and efficiency of multiple choice question and patient management problem item formats in assessment of clinical competence. Med Educ. 1985;19(3):238-47.

6. Berk RA. A Consumer's Guide to Multiple-Choice Item Formats That Measure Complex Cognitive Outcomes. pearson assessment and information. http://www. pearsonassessments. com/hai/images/NES_Publications/1996_12Berk_368_1. pdf. Accessed November; 2012.

7. Case SM, Swanson DB. Constructing written test questions for the basic and clinical sciences: National Board of Medical Examiners Philadelphia, PA; 1998.

8. Zoanetti N, Beaves M, Griffin P, Wallace EM. Fixed or mixed: a comparison of three, four and mixed-option multiple-choice tests in a Fetal Surveillance Education Program. BMC Med Educ. 2013;13:35.

9. Taylor AK. Violating Conventional Wisdom in Multiple Choice Test Construction. College Student Journal. 2005;39(1):141.

10. O'Mara D, Quinnell, R, Rothnie, I, et al; ExamBank: a pedagogic and administrative system to provide effective student feedback and stable assessment across disciplines. International Journal of Innovation in Science and Mathematics Education. In press.

11. AMSAC. Australian Medical Schools Assessment Collaboration: What do the differences mean? Unpublished paper presented to the annual ANZAHPE Conference Alice Spring NT June 2011. 2011.

12. De Champlain AF. A primer on classical test theory and item response theory for assessments in medical education. Med Educ. 2010;44(1):109-17.

13. M.H. WBDaS. Best Test Design. Rasch Measurement. Chicago: MESA Press; 1979.

14. SPSS I. IBM SPSS Statistics Version 21. Boston, Mass: International Business Machines Corp. 2012.

15. Bond TG, Fox CM. Applying the Rasch model: Fundamental measurement in the human sciences: Psychology Press; 2013.

16. Naeem N, van der Vleuten C, Alfaris EA. Faculty development on item writing substantially improves item quality. Advances in Health Sciences Education. 2012;17(3):369-76.

17. Dodds AE, Reid KJ, Conn JJ, Elliott SL, McColl GJ. Comparing the academic performance of graduate-and undergraduate-entry medical students. Med Educ. 2010;44(2):197-204.

1: AMSAC representation by medical school and medical students					
AMSAC representation	**Year of administration**				
	2009	**2010**	**2011**	**2012**	**2013**
Number of medical schools AMSAC	6	6	8	11	12
Proportion of all medical schools	32%	32%	42%	58%	63%
Number of students assessed	1035	1293	1666	2358	2377
Population for AMSAC schools[1]	1109	1321	1716	2383	2463
Proportion of all medical students in equivalent year*	33%	39%	51%	65%	68%

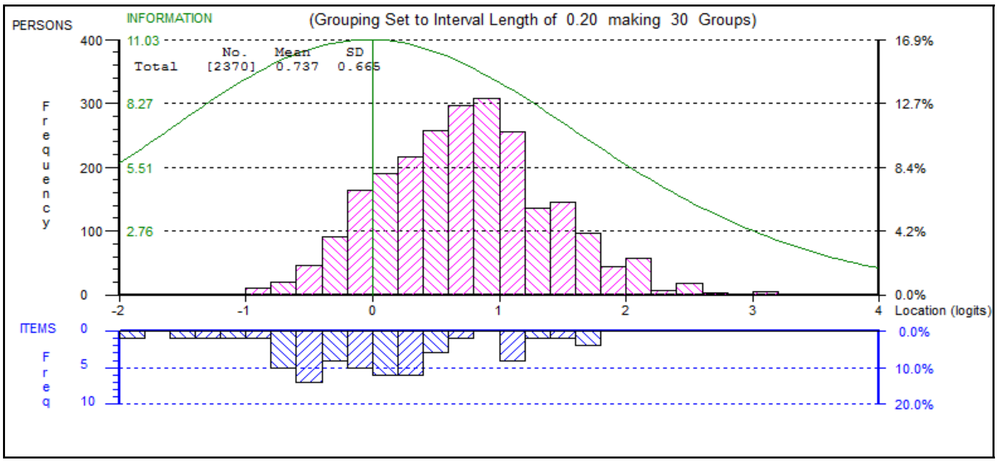
*Source: Medical Deans of Australia and New Zealand (MDANZ.) Table 2 (a): Total student enrolments 2013 by year of course (Australia): Medical Deans Australia and New Zealand; 2013. Available from: http://www.medicaldeans.org.au/wp-content/uploads/Website-Stats-2013-Table-2.pdf.

4: Mean Rasch scores by school type and implementation method mean (sd)[number of students]					
Year of implementation		**2011**	**2012**	**2013**	**Total**
School type	**Small school**	0.94 (0.82) [431]	0.67 (0.71) [420]	0.51 (0.64) [638]	0.68 (0.73) [1489]
	Large school	1.03 (0.64) [1235]	1.03 (0.86) [1938]	0.89 (0.70) [1739]	0.98 (0.75) [4912]
	School leaver entry	0.82 (0.70) [635]	0.61(0.68) [985]	0.74 (0.71) [775]	0.71 (0.70) [2395]
	Graduate entry	1.12 (0.66) [1031]	1.23 (0.85) [1373]	0.81 (0.69) [1602]	1.03 (0.77) [4006]
Implementation method	**Less than 90 seconds per MCQ**	0.92 (0.59) [556]	1.03 (0.95) [1219]	0.79 (0.73) [1205]	0.91 (0.81) [2980]
	90 seconds or more per MCQ	1.05 (0.74) [1110]	0.91 (0.70) [1139]	0.79 (0.67) [1172]	0.91 (0.71) [3421]
	Single assessment	1.24 (0.76) [575]	1.00 (0.72) [791]	0.68 (0.68) [588]	0.98 (0.75) [1954]
	More than one assessment	0.88 (0.62) [1091]	0.95 (0.90) [1567]	0.82 (0.71) [1789]	0.88 (0.76) [4447]
TOTAL		1.01 (0.69) [1666]	0.97 (0.84) [2358]	0.79 (0.70) [2377]	0.91 (0.76) [6401]

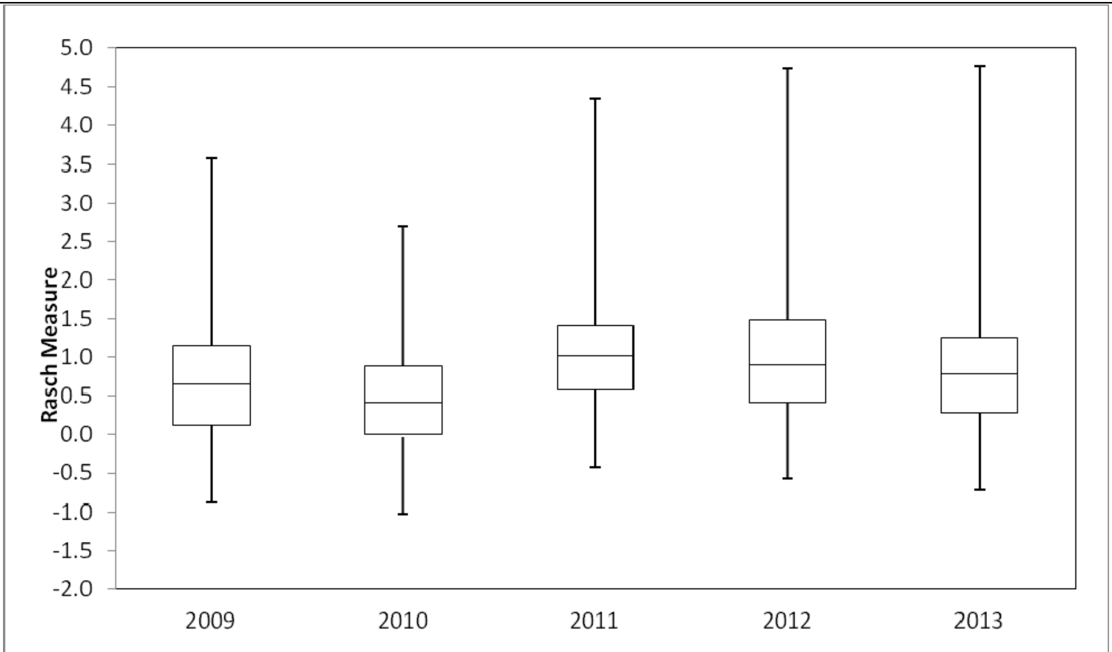
Photographs, graphs and illustrations

Photographs, graphs and illustrations may be inserted into this document for the purposes of submitting your article. If we decide to proceed with your article, you will need to provide separate high-quality versions of your photos and illustrations in appropriate file formats (see Instructions to authors) before your article can be accepted for publication.

2: 2013 AMSAC Student and Item Rasch distributions



3: AMSAC student performance over time



Each box displays the median score (horizontal line within the box) and the inter-quartile

APPENDICIES - Online at mja.com.au

Appendix 1: AMSAC Collaborators 2013

Australian National University
Deakin University
Joint Medical Program (JMP) for the Universities of Newcastle and New England.
Monash University
Sydney Medical School
The University of Adelaide
The University of Melbourne
The University of Notre Dame Freemantle
The University of Notre Dame Sydney
The University of Queensland
The University of Wollongong
University of Western Sydney

Appendix 2: AMSAC curriculum blueprint

Domain	Discipline and systems	Number of Questions
Function	Cellular systems (biochemistry, molecular biology, microbiology)	4
	Cardiovascular	3
	Endocrinology	3
	Gastrointestinal and Hepatology	2
	Haematology	2
	Immunology	2
	Neuroscience	3
	Pathology	1
	Pharmacology	2
	Renal	3
	Reproductive	2
	Respiratory	3
	Sub-total	**30**
Structure	Abdomen	3
	Lower limb	3
	Upper limb	3
	Head and Neck	2
	Nervous System	3
	Thorax	3
	Histology	3
	Sub-total	**20**
TOTAL		**50**

Appendix 3: AMSAC implementation variation

Number of questions used	2009	2010	2011	2012	2013
All 50 questions	3	3	3	2	1
40-49 used	2	2	3	5	7
25-39 used	0	0	2	1	4
Less than 25 used	1	1	1	3	1
Embedded in 1 exam	6	5	4	5	5
Embedded in >1 exams	0	1	4	6	7
Less than 90 seconds	3	3	2	5	5
90 seconds or more	3	3	6	6	7
TOTAL SCHOOLS	**6**	**6**	**8**	**11**	**12**

Appendix 4: ANOVA results by school type and implementation method

ANOVA Model by effects examined	Sum of Squares	df	F	p	Partial eta-squared
ANOVA 1 Total	174.4	5	63.3	<0.001	0.047*
AMSAC Year	63.1	2	57.2	<0.001	0.020*
Size of school	84.6	1	153.4	<0.001	0.020*
Year by Size of school	19.2	2	17.4	<0.001	0.010*
Error	3526.2	6395			
ANOVA 2 Total	315.5	5	119.2	<0.001	0.085**
AMSAC Year	39.1	2	36.9	<0.001	0.011*
Entry Type	161.0	1	304.2	<0.001	0.045*
Year by Entry type	79.5	2	75.1	<0.001	0.023*
Error	3385.1	6395			
ANOVA 3 Total	72.9	5	25.7	<0.001	0.020*
AMSAC Year	49.7	2	43.8	<0.001	0.014*
Time per MCQ	0.1	1	0.2	0.681	na
Year by Time per MCQ	14.8	2	13.0	<0.001	0.004
Error	3627.7	6395			
ANOVA 4 Total	117.7	5	42.0	<0.001	0.032*
AMSAC Year	85.7	2	76.5	<0.001	0.023*
Number of Assessments	10.7	1	19.0	<0.001	0.003
Year by Number of assessments	52.4	2	46.8	<0.001	0.014*
Error	3582.9	6395			

* Low effect size ** Moderate effect size